1-1-2014

Learning diagnostic diagrams in transport-based data-collection systems

Vu Tran
University of Wollongong, vtt921@uowmail.edu.au

Peter W. Eklund
University of Wollongong, peklund@uow.edu.au

Christopher David Cook
University of Wollongong, chris_cook@uow.edu.au

Learning diagnostic diagrams in transport-based data-collection systems

Abstract

Insights about service improvement in a transit network can be gained by studying transit service reliability. In this paper, a general procedure for constructing a transit service reliability diagnostic (Tsrd) diagram based on a Bayesian network is proposed to automatically build a behavioural model from Automatic Vehicle Location (AVL) and Automatic Passenger Counters (APC) data. Our purpose is to discover the variability of transit service attributes and their effects on traveller behaviour. A Tsrd diagram describes and helps to analyse factors affecting public transport by combining domain knowledge with statistical data.

Keywords

data, transport, systems, diagrams, collection, diagnostic, learning

Disciplines

Engineering | Science and Technology Studies

Publication Details

Tran, V. The., Eklund, P. & Cook, C. David. (2014). Learning diagnostic diagrams in transport-based data-collection systems. Lecture Notes in Computer Science, 8502 560-566.

Learning Diagnostic Diagrams in Transport-Based Data-Collection Systems

Vu The Tran, Peter Eklund*, and Chris Cook

Faculty of Engineering and Information Science
University of Wollongong
Northfields Avenue, NSW 2522, Australia

Abstract. Insights about service improvement in a transit network can be gained by studying transit service reliability. In this paper, a general procedure for constructing a transit service reliability diagnostic (TSRD) diagram based on a Bayesian network is proposed to automatically build a behavioural model from Automatic Vehicle Location (AVL) and Automatic Passenger Counters (APC) data. Our purpose is to discover the variability of transit service attributes and their effects on traveller behaviour. A TSRD diagram describes and helps to analyse factors affecting public transport by combining domain knowledge with statistical data.

Keywords: AI applications, knowledge discovery, Bayesian networks, transit service reliability

1 Introduction

> *"The transit industry is in the midst of a revolution from being data poor to data rich. Traditional analysis and decision support tools required little data, not because data has little value, but because traditional management methods had to accommodate a scarcity of data" [4].*

Automatic Vehicle Location (AVL) and Automatic Passenger Counters (APC) lead to big data and it is important to investigate how (and if) this can be used to improve transport service reliability. The growth of public transport databases facilitates new approaches to help characterize reliability and – by so doing – improve service planning and operational control. Among knowledge discovery techniques, Bayesian networks – a characterisation of probabilistic knowledge by a graphical diagram – provide a comprehensive method of representing relationships and influences among nodes. Bayesian networks are a fundamental technique in pattern recognition and machine classification [5],

Many studies induce Bayesian networks from data. Oniśko et al. [7] experiment with Bayesian network parameters from small data sets and use Noisy-OR gates to reduce the data requirements in learning conditional probabilities. Nadkarni et al. [6] describe a procedure for constructing Bayesian networks from

* On leave in 2014 to the IT University of Copenhagen, Denmark.

T. Andreasen et al. (Eds.): ISMIS 2014, LNAI 8502, pp. 560–566, 2014.

expert domain knowledge using causal mapping. Tungkasthan et al. [8] propose a practical framework for automating the construction of a diagnostic Bayesian network from WWW data sources. In that work, a SMILE (Structural Modeling, Interface, and Learning Engine) Web-based interface allow one to perform Bayesian network diagnosis through the Web.

As can be seen from the literature above, a wide variety of studies have use Bayesian networks for knowledge discovery, however employing Bayesian networks to analyze service reliability using data derived from AVL and APC sources for public transport is novel. To date, the transit industry has lacked a measure of service reliability measured in terms of its impact on customers because traditional measures cannot express how reliability impacts on passengers' perceptions [4]. Our paper focuses on an approach for constructing a transit service reliability diagnostic (TSRD) diagram based on a Bayesian network. A TSRD diagram has the ability to represent cause-effect relationships between transit factors and expresses how each factor will impact on others. A TSRD diagram can used in three ways: (i) as a guide for identifying the causes of service unreliability; (ii) as a learning component for real-time decision making and; (iii) as an offline analysis tool to improve service quality.

The remainder of the paper is organized as follows. The proposed methodology for constructing the TSRD diagram is presented in Section 2. The case study and experimental results are reported and discussed in Section 3.

2 Methodology

Service reliability in a public transport network can be considered as the variability of service attributes and their effect on passenger behaviour. A TSRD diagram based Bayesian network – a prediction-oriented method – is built to provide a better understanding of what causes problems in the transit system, prevent these problems through better service planning and operational management, and develop strategies to correct problems once they appear. A TSRD diagram is represented via a network $\mathcal{N}(\mathcal{G}, \Theta)$, where $\mathcal{G} =< \mathcal{U}, \mathcal{E} >$ is a directed acyclic graph, $\mathcal{U}$ is a set of nodes expressed as $\mathcal{U}\{u_1, u_2, ..., u_n\}$, $\mathcal{E}$ is a set of arcs, and Θ represents a set of conditional probability distributions.

Constructing a TSRD diagram involves of four steps: (1) preparation of transit discovery data set, (2) determining an initial TSRD diagram, (3) learn the TSRD structure and set parameters from training data, (4) assess/test the TSRD diagram. Since data from AVL and APC sources are heterogeneous and uncertain, the initial step combines data from various sources and tables into one dataset which can then be used in the discovery process. The second step is the construction of an initial TSRD diagram, based on cause-effect relationships to draw links between transit variables. The initial TSRD diagram reveals the qualitative relationships between variables in public transit systems. Next, the structure of the initial TSRD diagram and the parameters of variables need to be learned from the dataset. Learning the structure, causal relations, and parameters of variables – which reveal the quantitative relationships between variables – from

the dataset is important for an comprehensible and extensible TSRD diagram. The final steps is the assessment and validation of the candidate TSRD diagram.

2.1 Preparation of Transit Discovery Dataset

Step 1: From the original public transit data set variables relevant for the study are considered and selected. The raw AVL and APC data is stored in Tables with the schemes: Stops(StopName, Longitude, Latitude, SegmentID, StopNumber), Buses(BusID, Longitude, Latitude, Timestamp, Speed, SegmentID), Passengers(BusID, Longitude, Latitude, Timestamp, Counts, On/Off).

The raw data is normalized by combining, matching and processing data from the three tables to expose the variables required for analysis; this involves extraction and transformation of the attributes. This process is usually project-specific and the variables may vary, depending on how the TSRD diagram is to be used. In our case, all data is integrated into a single dataset including all of the attributes and their possible states that will be considered for the study of service reliability. Table 1 represents all combined attributes that are used.

Table 1. Description of attributes

No.	Variables	Possible states
1	vehicle Speed $\mathcal{V}$	{Slow, Normal, Fast}
2	vehicle position $\mathcal{X}$	{OnSchedule, OffSchedule}
3	running time $\mathcal{R}$	{OnTime, LessThan5MinLate, MajorLate}
4	passenger alighting $\mathcal{A}$	{Low, Normal, High}
5	passenger boarding $\mathcal{B}$	{Low, Normal, High}
6	dwell time $\mathcal{D}$	{Negligible, Major, Minor}
7	in-vehicle load $\mathcal{L}$	{Normal, Excessive, Unaccepted}
8	passenger wait time $\mathcal{T}_{wait}$	{Negligible, Major, Minor}
9	headway adherence $\mathcal{H}_{adherence}$	{Negligible, Major, Minor}
10	passenger comfort $\xi_{comfort}$	{Good, Accepted, Unaccepted}
11	service reliability $\mathcal{SR}$	{Yes, No}

2.2 Determining an Initial TSRD Diagram

Step 2: In this step we define the goals and understanding of what should be done with the TSRD diagram.

Question: "What causes headway irregularity?"
Answer: "Passengers alighting or boarding and the bus waiting for passengers running to catch the bus"
Modeling: Draw arcs from those nodes to the *Headway adherence* node.

After deciding what variables and states to model, an initial TSRD diagram N^0 is constructed by considering conditional independence by drawing causal links

among nodes following question and answer examples such as the one above. To establish the causal relationships, it is helpful to ask direct questions about dependence between variables. Once identified, arcs are added from those causal variables to the affected variable. Probabilities on the edges are obtained initially by subjective estimates. First diagram to the left in Fig. 1 depicts the TSRD diagram N^0 by combining knowledge of bus operations and asking cause-effect questions. The main interest of service reliability diagnosis is to identify causes of unreliability. The context variables in this case are the background information about passengers alighting, passengers boarding, bus position and speed.

2.3 Learning Structure and Parameters from Data Set

Step 3: The initial TSRD diagram is often not good enough because there is often not enough causal knowledge to establish the full topology of the network model. Learning the full structure, causal relations and parameters from a data set are essential for refining and conditioning the TSRD diagram.

The applied structural Expectation Maximization (EM) algorithm requires an initial TSRD diagram N^0 and a dataset D as a starting point for iteration. Learning the probabilities of attributes of the TSRD diagram from data is a form of unsupervised learning. The objective here is to deduce a network that best describes the probability distribution over the training data $\mathcal{D}$.

The structural EM algorithm is an extension to the standard Expectation Maximisation algorithm [1] and is described in [2] and [3]. The algorithm performs a search in the joint space of structure and parameters. At each step, it can either find better parameters for the current structure, or select a new structure. The function Q is the expected score, given by:

Algorithm 1. Learning structure and parameters from dataset

input : $D = \{x^1, ..., x^n\}$: a data set
input : $N^0 = (\mathcal{G}', \Theta^0)$: an initial network
output: $N^* = (\mathcal{G}^*, \Theta^*)$: return the candidate network

begin
  Loop for n= 1,2,... until convergence **begin**
    Find a model G^{n+1} that maximises $Q(\mathcal{G}, \Theta : \mathcal{G}^n, \Theta^n)$
    Let $\Theta^{n+1} = Q(\mathcal{G}^{n+1}, \Theta : \mathcal{G}^n, \Theta^n)$
  return N^*

$$Q(\mathcal{G}, \Theta : \mathcal{G}^*, \Theta^*) = E[\log P(O, h : \mathcal{G}, \Theta) - \text{Penalty}(\mathcal{G}, \Theta)] \quad (1)$$

where O are the observed variables, h are the values of the hidden variables, and the penalty depends on the dimensionality of $\mathcal{G}$. The procedure converges to a *local* maxima.

2.4 Assess Structure and Parameters

Step 4: Crucial to the methodology is that the structure and parameters of the model are validated. The structure evaluation reveals if important variables

have been overlooked, if irrelevant nodes have been included, or if node values are inappropriate. Validation confirms that the model is an accurate representation of the domain. The evaluation and validation consists, in this case, of comparing the behaviour of a network with expert judgements.

3 Results and Discussion

A case study of bus operations on the Gwynneville-Keiraville bus route in the regional city of Wollongong, Australia – population approx. 300,000 – is used to demonstrate and test the proposed method. AVL and APC units are installed for buses on the UniShuttle service to capture the data. As at the end of July 2012 there were a total of 1,844,964 vehicle (bus) location events stored in a MySQL database on our servers. The average monthly number of vehicle events captured is 132,000. There are an average of 4,630 passenger events per month captured and this number will increase 1000% when APC devices are installed on all buses in the fleet as is proposed.

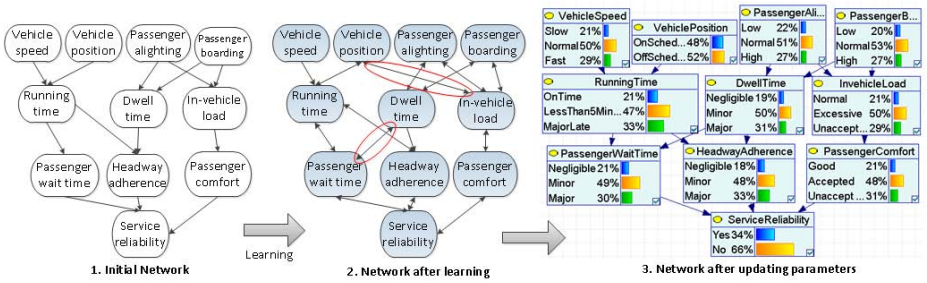

Fig. 1. TSRD diagram: construction process

In diagram centre of Fig. 1, after learning, the topology is modified, with two new arcs added. The new connections are from *VehiclePosition* node to *InvehicleLoad* node and from the *DwellTime* to *PassengerWaitTime* node. After the validation step, the casual connection from *VehiclePosition* to *InvehicleLoad* nodes is eliminated as the expert judgement is that it is inappropriate. In diagram to right of Fig. 1, the conditional probability tables (CPTs) annotate the nodes. These represent how much reliability exists in the current transit network data. Each row in a CPT contains the conditional probability of each node value for a conditioning case.

Based on our results, measures to reduce service unreliability should balance passenger wait time, passenger comfort and headway adherence as the service unreliability is similarly impacted (posterior probability) by each of these three indicators. Transport management would be advised to better control these figures so that service reliability is improved. Of the indirect factors, dwell time has the greatest posterior probability 0.81 (minor and major), as these factors

affect passenger wait time and headway adherence. The posterior probability for in-vehicle load and running time is 0.79. These probabilities are high enough to indicate that transportation management should pay more attention to scheduling and planning to improve running time and reduce passenger load. This is a reassuring recommendation that validates the model: namely that service reliability is improved by more buses and fewer passengers.

The use of TSRD diagram represents three aspects: Filtering, Smoothing, and Learning component. Filtering of TSRD diagram is used to compute the belief state of the posterior distribution of transit service reliability over the most recent state, given all the observations (evidence) of the public transit factors made so far. Smoothing of the TSRD diagram is carried out to compute the posterior distribution over a past state, given all the observations (evidence) of the public transit factors made to date. The TSRD diagram is used as a knowledge representation of prior knowledge of real-time control strategies. Learning enables the transit systems to function in initially unfamiliar environments and to become more competent over time than its preliminary knowledge state.

4 Conclusion

Modern scheduling and customer service monitoring is oriented around extreme values (outrider events) rather than traditional mean values. This is mainly because of the large sample sizes produced by automatic data collection and so attention focuses on unusual events.

As these kinds of information are characterized as heterogeneous and uncertain, a TSRD diagram based Bayesian networks is presented in this paper to serve as our knowledge model to analyze automatic data collection. The TSRD diagram has the advantages of an intuitive visual representation with a sound mathematical basis in Bayesian probability and provides an effective approach for analysis of public transit systems to reveal the hidden structure and its relationships, and more importantly, its rules. A case study is used to evaluate and demonstrate the use of TSRD diagram.

References

1. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 1–38 (1977)
2. Friedman, N.: Learning belief networks in the presence of missing values and hidden variables. In: Machine Learning International Worshop then Conference, pp. 125–133. Morgan Kaufmann Publishers, Inc. (1997)
3. Friedman, N.: The bayesian structural em algorithm. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, pp. 129–138. Morgan Kaufmann Publishers Inc. (1998)
4. Furth, P.G.: Using archived AVL-APC data to improve transit performance and management. Transportation Research Board National Research, vol. 113 (2006)
5. Korb, K.B., Nicholson, A.E.: Bayesian artificial intelligence, vol. 1. CRC Press (2004)

6. Nadkarni, S., Shenoy, P.P.: A causal mapping approach to constructing bayesian networks. Decision Support Systems 38(2), 259–281 (2004)
7. Oniśko, A., Druzdzel, M.J., Wasyluk, H.: Learning bayesian network parameters from small data sets: Application of noisy-or gates. International Journal of Approximate Reasoning 27(2), 165–182 (2001)
8. Tungkasthan, A., Jongsawat, N., Poompuang, P., Intarasema, S., Premchaiswadi, W.: Automatically building diagnostic bayesian networks from on-line data sources and the smile web-based interface. In: Jao, C.S. (ed.) Decision Support Systems, pp. 321–334 (2010)